Untersuchungen zur Ursache unterschiedlicher Ergebnisse sehr ähnlicher Viktimisierungssurveys

Schnell, Rainer; Kreuter, Frauke

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

UNTERSUCHUNGEN ZUR URSACHE UNTERSCHIEDLICHER ERGEBNISSE SEHR ÄHNLICHER VIKTIMISIERUNGSSURVEYS*

Rainer Schnell und Frauke Kreuter

Zusammenfassung: Zwei 1997 in der Bundesrepublik Deutschland erhobene sehr ähnliche Viktimisierungssurveys (gleiche Grundgesamtheit, gleiches Institut, fast gleiches Design, teilweise identische Interviewerstäbe, identische Operationalisierung etc.) zeigten so deutliche Unterschiede in den geschätzten Opferraten, dass selbst um Klumpeneffekte und Designgewichte korrigierte Konfidenzintervalle signifikante Unterschiede zwischen den Surveys indizieren. Zur Erklärung dieser Unterschiede werden eine Reihe möglicher methodischer Details der beiden Surveys kritisch diskutiert und in ihren Konsequenzen quantitativ abgeschätzt. Als vermutliche Ursache für die unterschiedlichen Survey-Ergebnisse werden die Rahmenbedingungen der Arbeit der Interviewer identifiziert. Das Ergebnis zeigt eindringlich die Wichtigkeit solcher kaum je dokumentierten und noch seltener kontrollierten Details der Feldarbeit bei sozialwissenschaftlichen Erhebungen.

I. Einleitung

Die Gültigkeit von Survey-Ergebnissen ist von großer praktischer und theoretischer Bedeutung. Dies gilt in Forschungsgebieten mit unmittelbarer politischer Relevanz wie z.B. für Opferbefragungen, die in der kriminalpolitischen Diskussion der Bundesrepublik eine zunehmende Bedeutung gewinnen, in besonderem Maße. Sollen Veränderungen in berichteten Opferraten in Viktimisierungssurveys als Indikatoren für den Erfolg oder Misserfolg politischer Maßnahmen oder als Indikator für sozialen Wandel interpretiert werden, dann muss ausgeschlossen werden können, dass „signifikante" Unterschiede zwischen den Ergebnissen verschiedener Surveys die Folge kaum je berichteter methodischer Durchführungsdetails sind (Turner 1984). Um die Effekte solcher Details entdecken zu können, benötigt man eine oder mehrere nahezu vollständige Replikationen eines Surveys. Zwei 1997 durchgeführte Erhebungen bieten die in der Bundesrepublik außergewöhnliche Gelegenheit, „signifikante" Unterschiede zwischen zwei sehr ähnlichen Surveys exemplarisch zu untersuchen.

* Wir sind den Diskutanten im Anschluss an zwei Vorträge vorläufiger Versionen dieses Artikels (Methodensektion der DGS in Mannheim und Südwestdeutsches Kriminologisches Kolloquium in Konstanz, beide im Sommer 1999) sowie Margarete Boos, Günther Heller, Johannes Kopp, Jürgen Lass und dem anonymen Gutachter für kritische Kommentare dankbar.

II. Problemstellung und Datenbasis

Das im Auftrag des Bundesministeriums für Justiz 1997 durchgeführte Projekt einer Opferumfrage bestand aus zwei Surveys.[1] Eine 20.070 Befragte umfassende Mehrthemenumfrage (MTU) enthielt Fragen nach Opfererfahrung, Anzeigeerstattung und Gründe für Nichtanzeige. Ergänzend dazu wurde eine kleinere Befragung im Rahmen des Sozialwissenschaften-Bus III/97 (SWB) mit 3.272 Befragten durchgeführt, die zusätzliche Fragen zum „Subjektiven Sicherheitsgefühl" und „Social Disorder" enthielt.

Die in den beiden Studien ermittelten Opferraten unterscheiden sich erheblich.[2] Die Opferrate liegt in der MTU bei 15,9 Prozent, in der SWB bei 19,5 Prozent (vgl. *Tabelle 1*).

Tabelle 1: Exakte Binomial-Konfidenzintervalle der beiden Studien

Survey	Fälle	Mittelwert Opferwerdung	Standard-fehler	95%-Konfidenzintervall
MTU	20070	0.1587	0.0026	0.1537 – 0.1639
SWB	3272	0.1950	0.0069	0.1815 – 0.2090

Der Unterschied in den Opferraten dieser beiden Studien ist insbesondere deshalb bedeutsam, da sich die zugehörigen Konfidenzintervalle nicht überlappen.[3] Das 95%-Konfidenzintervall in der MTU reicht von 15,4 Prozent bis 16,4 Prozent, im SWB von 18,2 bis 20,9 Prozent.[4] Der Unterschied zwischen den Stichproben ist damit signifikant. Der deutliche Unterschied zwischen den Stichproben ist offensichtlich erklärungsbedürftig. Im Folgenden sollen mögliche Erklärungen für diesen Unterschied entwickelt und – soweit möglich – getestet werden. Dabei wird zwischen Erklärungen

1 Wir danken dem beteiligten Institut, dem Bundesministerium für Justiz sowie der Arbeitsgruppe „Strafrechtliche Rechtstatsachenforschung und empirische Kriminologie" der Universität Konstanz unter der Leitung von Wolfgang Heinz für den Zugang zu diesen Datensätzen.

2 Als Opfer wurden diejenigen Befragten kategorisiert, die angaben, Opfer mindestens eines der folgenden Verbrechen geworden zu sein: PKW-Diebstahl, Fahrrad-Diebstahl, PKW-Teildiebstahl, PKW beschädigt, Eigentum beschädigt, Einbruchsversuch, Gewaltandrohung, Taschendiebstahl, freches Benehmen, tätlicher Angriff, Betrug (vgl. Heinz und Spieß 1995: 101).

3 Ein Konfidenzintervall versucht das Ausmaß der Unsicherheit bei der Schätzung eines Parameters zu quantifizieren. Zwar lässt sich für ein gegebenes 95%-Konfidenzintervall keine Wahrscheinlichkeit angeben, ob es den Parameter enthält oder nicht, aber würde man einhundert Stichproben aus der Grundgesamtheit ziehen, so würden ca. 95% der berechneten Konfidenzintervalle den Parameter enthalten. Überlappen sich die Konfidenzintervalle zweier verschiedener Stichproben nicht, so stammen die Stichproben mit hoher Wahrscheinlichkeit nicht aus der gleichen Grundgesamtheit.

4 Mit Hilfe des Statistikpaketes *Stata* (Stata Corp. 1997) wurden für die binäre Variable Opfer (0,1) der Mittelwert und die Konfidenzintervalle für binomialverteilte Variablen mit der Clopper-Pearson-Methode (Agresti und Coull 1998: 119) berechnet.

durch üblicherweise berichtete Erhebungsdetails und Erklärungen durch üblicherweise nicht berichtete Erhebungsdetails unterschieden.

III. Üblicherweise berichtete Erhebungsdetails als mögliche Ursachen

1. Definition der Grundgesamtheit

Ein signifikanter Unterschied zwischen zwei Stichproben indiziert in der Regel, dass die beiden Stichproben nicht aus der gleichen Grundgesamtheit stammen. Der einzige Unterschied in der Definition der Grundgesamtheit der beiden Erhebungen besteht aber in der Berücksichtigung der 16- bis 18-Jährigen in der MTU; ansonsten verwenden beide Studien die in der BRD übliche Populationsdefinition.[5] Schließt man die unter 18-Jährigen aus der MTU aus, führt dies sogar zu einer kleinen Vergrößerung des Unterschiedes zwischen den beiden Stichproben um 0,2 Prozent (vgl. *Tabelle 2*). Die unterschiedliche Definition der Grundgesamtheit ist also nicht die Erklärung für den signifikanten Unterschied zwischen den Stichproben.

Tabelle 2: Exakte Binomial-Konfidenzintervalle der beiden Studien (nur Volljährige)

Survey	Fälle	Mittelwert Opferwerdung	Standard-fehler	95%-Konfidenzintervall
MTU	19682	0.1567	0.0026	0.1517 – 0.1619
SWB	3272	0.1950	0.0069	0.1815 – 0.2090

2. Auswahlverfahren

Selbst bei identischer Grundgesamtheitsdefinition können Unterschiede im Auswahlverfahren zu unterschiedlichen Ergebnissen führen. Beide Erhebungen verwenden ein dreistufiges Auswahlverfahren in Anlehnung an das ADM-Design (vgl. Schnell, Hill und Esser 1999: 268–269).

In der ersten Auswahlstufe wurde in der MTU eine bundesweite flächendeckende Stichprobe von insgesamt n = 2.100 Sampling-Points gezogen. Für den Sozialwissenschaften-Bus wurden 630 verschiedene Sampling-Points ausgewählt.[6] Das Auswahlver-

5 „... alle Personen mit Wohnsitz in der Bundesrepublik Deutschland, die ... in Privathaushalten der Bundesrepublik Deutschland lebten. Die Grundgesamtheit schloß die deutsche Wohnbevölkerung ein sowie die ausländische Wohnbevölkerung, soweit die Möglichkeit bestand, das Interview in deutscher Sprache durchzuführen" (Methodendokumentation 1998: 2). Diese Definition der allgemeinen Bevölkerung ist nicht unproblematisch, bestimmte Bevölkerungsteile werden hierdurch systematisch ausgeschlossen, z.B. Bewohner von Altersheimen, Nichtsesshafte und Anstaltsbewohner (vgl. Schnell 1991). Solche Populationen unterscheiden sich von der übrigen Bevölkerung deutlich in ihrer Viktimisierungsrate (vgl. Lynn 1997).

6 Sowohl die Zuordnung dieser Points zu Netzen als auch die Auswahl der Netze wurde im

fahren der Points im SWB sollte zu einer zu Gunsten der neuen Bundesländer disproportional geschichteten Stichprobe führen. Da die Opferrate in den neuen Bundesländern höher liegt, sollte die Berücksichtigung der Designgewichte zu einer Absenkung der Opferrate im gesamten Datensatz führen. Überraschenderweise bleibt die Gesamtopferrate fast konstant, gleichgültig welches der verschiedenen Gewichte verwendet wird.[7]

Die zweite Auswahlstufe bestand in der MTU aus folgender Variante eines „Random Walks": Die Interviewer erhielten für die ihnen zugewiesenen Sampling-Points eine Startadresse. Ausgehend von dieser sollte nach einem vorgegebenen Wegeschema jeder dritte Haushalt kontaktiert werden. Handelt es sich um einen Haushalt der definierten Grundgesamtheit, hatten sich die Interviewer um „die Realisierung eines Interviews zu bemühen" (Methodendokumentation 1998: 5).

Im SWB bestand die zweite Auswahlstufe aus einer anderen Variante des „Random Walks": Ausgehend von einer Startadresse wurde jeder dritte Privathaushalt auf einem „Adressenauflistungsblatt" notiert, bei dem das Institut zuvor acht bzw. neun Adressenfelder markierte. Die in diesen Feldern eingetragenen Adressen bilden die Brutto-Ausgangsstichprobe.

Die dritte Auswahlstufe im ADM-Design besteht aus der Auswahl der Zielpersonen aus dem Haushalt. In der MTU erfolgte diese Auswahl durch die Befragung derjenigen Person, die als letzte Geburtstag hatte. Im SWB erfolgte die Auswahl anhand eines Zufallszahlenauswahlschlüssels.

Nimmt man an, dass die Auswahlverfahren fehlerfrei durch das Institut und die Interviewer durchgeführt wurden, so gibt es keinen theoretischen Grund, warum diese unterschiedlichen Auswahlverfahren zu unterschiedlichen Opferraten in den beiden Stichproben führen sollten.

Allerdings zeigt diese Darstellung der Details des Auswahlverfahrens, dass es sich bei beiden Stichproben – wie fast immer in der Praxis der empirischen Sozialforschung – nicht um einfache Zufallsstichproben, sondern um komplexe Bevölkerungsstichproben handelt. Aus diesem Grund sind die üblicherweise – wie oben – berechneten Konfidenzintervalle irreführend. Die tatsächlichen Konfidenzintervalle sind fast immer wesentlich größer. Dies soll kurz erläutert werden.

Exkurs: Konfidenzintervalle in komplexen Bevölkerungsstichproben

Die Größe der Konfidenzintervalle hängt vor allem von den geschätzten Standardfehlern ab. Die meisten einführenden Lehrbücher der Statistik geben für die Schätzung der Standardfehler lediglich Formeln an, die auf der Annahme einfacher Zufallsstich-

---

Methodenbericht des Instituts nicht eindeutig beschrieben bzw. die Beschreibung steht im Gegensatz zu den Angaben im Datensatz.

7 Der Datensatz enthält mehr als eine Gewichtungsvariable. Leider geht aus der Methodendokumentation nicht hervor, wie diese Faktoren berechnet wurden. Weiterhin sind die Gewichtungsfaktoren für den MTU nicht getrennt für die über 18-Jährigen ausgewiesen, so dass Berechnungen mit dem gewichteten Datensatz, bei dem die unter 18-Jährigen ausgeschlossen wurden, streng genommen nicht korrekt sind.

proben basieren. In der Forschungspraxis sind hingegen komplexere Stichprobendesigns üblich. Diese erfordern andere Schätzformeln, da bei komplexeren Stichproben die Standardfehler durch die auf einfachen Zufallsstichproben basierenden Formeln unterschätzt werden. Da es sich bei beiden Stichproben um mehrstufige Klumpenauswahlverfahren handelt, müssen für einen korrekten Vergleich der Konfidenzintervalle andere Schätzer der Standardfehler verwendet werden.

Bei den in der BRD üblichen Klumpenstichproben werden z.B. in der ersten Stufe als Klumpen die Sampling-Points ausgewählt und in der zweiten Stufe innerhalb dieser Klumpen die Untersuchungseinheiten (Haushalte) gezogen. Diese Art der Ziehung hat deutliche Kosten- und Organisationsvorteile.

Nun besteht die Möglichkeit, dass die Elemente innerhalb eines Klumpens homogener sind als zufällig aus der Grundgesamtheit ausgewählte Elemente. Dies kann mehrere Ursachen haben. Zum einen können die Elemente auf Grund ihrer räumlichen Nähe homogener sein als zufällig aus der Grundgesamtheit ausgewählte Elemente. In Hinsicht auf die Opferraten bedeutet dies, dass Personen, die in einem Sampling-Point wohnen, ähnliche Opferraten besitzen. Zum anderen kann der Datenerhebungsprozess zu einer Homogenisierung der Elemente eines Klumpens beitragen. Dies kann zum Beispiel durch den Einfluss des Interviewers auf den Erhebungsprozess geschehen.[8] Der homogenisierende Einfluss des Interviewers kann sich bei der Auswahl der Erhebungseinheiten zeigen, aber auch in der Interviewsituation selbst.

Bei der Auswahl der Erhebungseinheiten könnten Interviewer z.B. lediglich kooperative Personen befragen. Sollte die Kooperationsbereitschaft mit der Opferwerdung korreliert sein (z.B. weil Personen, die schon einmal Opfer geworden sind, weniger bereit sein könnten, Fremde in ihre Wohnung zu lassen), so wären die Opferraten der realisierten Interviews dieser Interviewer untereinander korreliert. Während der eigentlichen Datenerhebung kann der Interviewer durch die Interpretation der Fragen und die Protokollierung der Antworten homogenisierend wirken. In diesen Studien könnte dies z.B. dadurch erfolgen, dass der Interviewer Fragen, die ihm peinlich sein könnten (wie etwa Fragen nach sexueller Belästigung) umformuliert oder gänzlich meidet. Auch in diesem Fall werden die Opferraten bei den Interviews dieses Interviewers untereinander korreliert sein.

Zur Berechnung korrekter Konfidenzintervalle trotz des Effekts homogener Klumpen stehen eine Reihe verschiedener statistischer Verfahren zur Verfügung. Das einfachste Modell basiert auf der Berechnung von „Design-Effekten".[9]

Hierbei wird zunächst die Homogenität innerhalb der Cluster berechnet.[10] Diese Homogenität wird mit dem Intraklassenkorrelationskoeffizient *roh* (*r*ate *o*f *h*omogeneity) wiedergegeben (Kish 1965: 163).[11] Ausgehend von *roh* kann der „Design-Effekt" *deff* berechnet werden

8 Eine ausführliche Dokumentation dieser Prozesse findet sich bei Schnell (1997).

9 Auf die anderen Modelle wird hier nicht weiter eingegangen, zu Einzelheiten vgl. Schnell (1997).

10 Die Bezeichnung Cluster wird für die betrachtete Einheit verwendet, dies können alle Befragten innerhalb eines Klumpens (Sampling-Points) oder eines Interviewers sein.

11 *Roh* kann zum Beispiel mit einer einfachen Varianzanalyse berechnet werden:

$$\text{roh} = \frac{(F-1)b/n}{1+(F-1)b/n},$$

$$deff = 1 + (w - 1)roh$$

wobei w die durchschnittliche Clustergröße ist.[12] Die Wurzel aus *deff* wird als *deft* bezeichnet. Die um den Designeffekt korrigierten Konfidenzintervalle sind um diesen Faktor *deft* größer als die entsprechenden Konfidenzintervalle einer einfachen Zufallsstichprobe.

Bei den beiden vorliegenden Studien lassen sich die auf die Sampling-Points und Interviewer zurückgehenden Effekte auf Grund des Studiendesigns nicht voneinander trennen. Die Opferraten werden deshalb einmal nur unter Berücksichtigung der Sampling-Points als clusterdefinierende Variable und ein zweites Mal nur unter Berücksichtigung der Interviewer als clusterdefinierende Variable berechnet.

Durch die Berücksichtigung des homogenisierenden Einflusses der Sampling-Points können sich die Opferraten natürlich nicht verändern, wohl aber die Konfidenzintervalle.[13] Diese werden unter Berücksichtigung der Design-Effekte etwas größer („cluster = P" in *Abbildung 1*): In der MTU ergibt sich ein geschätztes Intervall von 0.149 bis 0.165, im SWB liegt das Konfidenzintervall zwischen 0.177 und 0.213. Verwendet man an Stelle des Points den Interviewer als Cluster, so vergrößern sich die Konfidenzintervalle nochmals („cluster = I" in *Abbildung 1*): Für die MTU liegt das Intervall zwischen 0.148 und 0.166 und für den SWB zwischen 0.172 und 0.218. Die Design-Effekte, die auf die Interviewer zurückgehen, liegen deutlich über den Effekten der Sampling-Points.[14] Berücksichtigt man neben den Design-Effekten noch die Gewichtungsfaktoren („cluster = Iw" in *Abbildung 1*), so werden die Konfidenzintervalle noch etwas größer (MTU: 0.150 bis 0.168; SWB: 0.170 bis 0.223). Aber auch diese Kombination führt nicht zu sich überlappenden Konfidenzintervallen: Der Unterschied zwischen den beiden Stichproben bleibt auch bei korrigierten Konfidenzintervallen signifikant (vgl. *Abbildung 1*). Es scheinen tatsächliche Unterschiede zwischen den beiden Surveys zu bestehen.

---

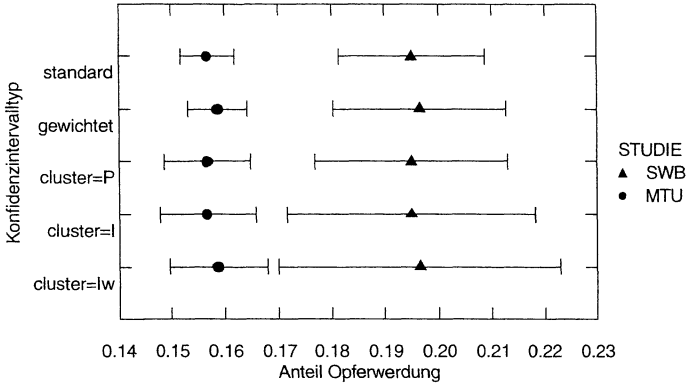
wobei F die F-Statistik aus der ANOVA Tabelle, k die Zahl der Cluster und n die Zahl der Fälle ist (Stata 1997: 381). *Roh* erreicht den Wert 1.0, falls alle Elemente innerhalb der Cluster den gleichen Wert annehmen. Die empirischen Werte für *roh* liegen üblicherweise bei 0.01 bis 0.02 (Groves 1989: 318).

12 Die Verhältniszahl *deff* dient dem Vergleich der in einem komplexen Survey berechneten Varianz und der Varianz, die unter der Bedingung einer einfachen Zufallsauswahl erreicht worden wäre. Ein *deff* von 1.0 indiziert die Abwesenheit eines Designeffekts.

13 Zu Gunsten einer leichteren Verständlichkeit wird mit den Mittelwerten der dichotomen Opfervariable gerechnet. Korrekt wäre eine Berechnung der Anteilswerte. Die Ergebnisse unterscheiden sich jedoch erst in der fünften Nachkommastelle. Verwendet wurde die Prozedur *svymean* der Version 6.0 des Statistikpaketes *Stata*.

14 Dass die Design-Effekte in der MTU höher als im SWB liegen, ist zum größten Teil durch die höhere Anzahl von Interviews pro Interviewer zu erklären.

Abbildung 1: Konfidenzintervalle des Anteils der Opferwerdung



Erläuterung zur Abbildung:
Konfidenzintervalltyp
„standard": Standard-Konfidenzintervall
„gewichtet": Standard-Konfidenzintervall mit gewichteten Daten
„cluster = P": Konfidenzintervalle mit Sampling-Point als clusterdefinierende Variable
„cluster = I": Konfidenzintervalle mit Interviewer als clusterdefinierende Variable „cluster = Iw": Konfidenzintervalle mit Interviewer als clusterdefinierende Variable und gewichteten Daten

3. Erhebungszeitpunkt

Es liegt nahe, den Unterschied zwischen den Stichproben durch eine zeitliche Veränderung der tatsächlichen Prävalenzraten zu erklären. Diese Erklärung scheitert aber daran, dass beide Erhebungen fast zeitgleich stattfanden.[15]

4. Operationalisierung

Die nicht weniger nahe liegende Erklärung durch unterschiedliche Operationalisierung scheidet bei diesen Erhebungen ebenfalls aus: Die Operationalisierungen der hier interessierenden Variablen sind vollkommen identisch.

5. Erhebungsinstrumente

Trotz identischer Operationalisierung der Opferwerdung weisen die Fragebögen erhebliche Unterschiede auf. Beide Erhebungen sind Omnibus-Studien, das heißt, dass neben den Fragen zur Opferwerdung Fragen zu einer ganzen Reihe anderer Themen ge-

15 Die MTU fand zwischen dem 6. Oktober 1997 und dem 9. Januar 1998 statt. Die Erhebung im Rahmen des SWB fand in der Zeit vom 30.10.1997 bis 31.12.1997 statt.

stellt wurden. Durch die unterschiedlichen Fragekontexte in beiden Erhebungsinstrumenten könnte der Unterschied zwischen den Stichproben bedingt sein.[16] Zwei Mechanismen erscheinen möglich: Zum einen könnten die Befragten Viktimisierungsfragen in einem Marktforschungskontext weniger ernst nehmen als in einem rein sozialwissenschaftlichen Fragekontext, zum anderen könnten Unterschiede in der Art der vorhergehenden Fragen zu verändertem Antwortverhalten führen.

Es gab mehrere Split-Versionen der Fragebögen, sowohl in der MTU als auch im SWB. In der uns zugänglichen Version der MTU kamen die Opferwerdungsfragen u.a. nach 17 Fragen zur Wohnung und 48 Fragen zu Kaffee und Schokolade, wobei die genauen Zahlen je nach Filterführung variieren können.

Im SWB wurden zunächst 53 Fragen zur Wohnsituation und 8 Arbeitsplatzfragen gestellt, dann kamen neun Fragen zur Kriminalitätsfurcht. Diesen folgten 31 Fragen zu Sprachgebrauch und Dialekt, 26 Fragen zu Kindern und Schule sowie drei Fragen zu politischen Einstellungen. Den Abschluss dieses Blocks bildeten die Opferwerdungsfragen.

Damit wären für beide Mechanismen die Anfangsbedingungen erfüllt: Bei der MTU war für die Befragten der Marktforschungskontext vor den Opferwerdungsfragen klar erkennbar, weiterhin wurden im SWB vor den Opferwerdungsfragen Fragen zur Kriminalitätsfurcht gestellt. Sollte die Fragereihenfolge „Furcht-Opferwerdung" die Erinnerung an Opfersituationen erleichtern, könnte die berichtete höhere Opferrate im SWB dadurch erklärt werden. Allerdings kann mit den vorliegenden Daten keine dieser Hypothesen direkt geprüft werden. Auf der Basis des bisherigen Wissens über Omnibus-Surveys scheint ein Effekt dieses Surveytyps zwar nicht ausgeschlossen, andererseits wurde bislang ein solcher Effekt auch nicht berichtet.[17] Angesichts der starken Verbreitung von Omnibus-Surveys besteht hier dringend Forschungsbedarf. Auch die Hypothese des möglichen Effekts der Fragereihenfolge „Furcht-Opfer" vs. „Opfer-Furcht" wurde bislang in der Literatur weder explizit formuliert noch wurde ein Test dieser Hypothese veröffentlicht.

Um den möglichen Effekt der Fragereihenfolge zu quantifizieren, wurden für diese Untersuchung der Unterschiede zweier Surveys zwei Experimente an der Universität Konstanz durchgeführt. In einem ersten Experiment wurden die Fragen in zwei in der Reihenfolge der Fragen verschiedenen Fragebogenformen („Furcht-Opfer", „Opfer-Furcht") im Sommersemester 1999 126 Studenten einer einführenden juristischen Vorlesung schriftlich vorgelegt. Es zeigte sich eine Prozentsatzdifferenz von –12,53 Prozent (44,23 vs. 56,76 Prozent).[18]

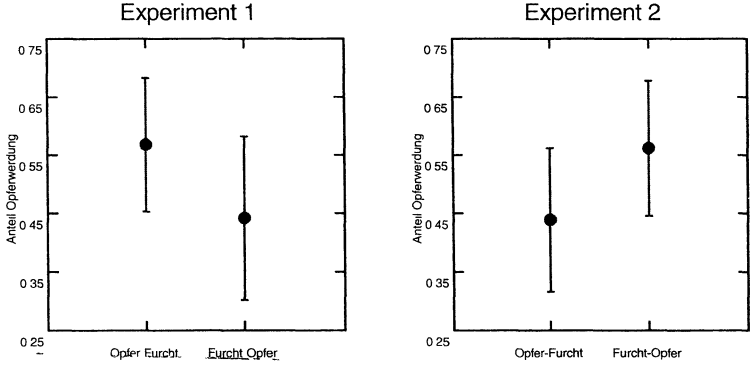
16 Zunächst lagen die vollständigen Erhebungsunterlagen (wie in der BRD bei Omnibus-Befragungen üblich) dem Auftraggeber nicht vor. Auf besondere Nachfrage wurde ausnahmsweise Einblick in die vollständigen Fragebögen gewährt.

17 Die Datenbanken *Sociofile, Solis* und *SRM* enthalten unter den Stichworten „Omnibus" bzw. „Mehrthemen*" keine Hinweise auf Tests des potenziellen Effekts eines Omnibus-Surveys gegenüber einem „single-purpose-survey" auf das Antwortverhalten bei identischen Fragen.

18 Diese Opferraten liegen deutlich höher als die Opferraten im SWB bzw. in der MTU. Betrachtet man jedoch in den beiden Erhebungen nur Personen zwischen 18 und 23 mit Fachhochschulreife oder Abitur, so liegt der Opferanteil bei 34 Prozent (SWB) bzw. 30

Dieser Effekt ist auf dem 5%-Niveau nicht signifikant von Null unterschieden (p = .17, Power = .22 zweiseitig bzw. Power = .33 einseitig bei p = .05). Um den möglichen Effekt des Studienfaches zu kontrollieren, wurde im Wintersemester 1999 ein zweites Experiment mit den 139 Erstsemestern einer einführenden Methodenveranstaltung (in der zweiten Semesterwoche) durchgeführt. In dieser Studie ergab sich eine Prozentsatzdifferenz von +12,22 Prozent (56,16 vs. 43,94 Prozent, vgl. *Abbildung 2*). Auch in dieser Studie ist kein von Null verschiedener Effekt der Fragereihenfolge feststellbar (p = .15, Power = .24 zweiseitig bzw. Power = 0.35 einseitig bei p = 0.05). Beide Ergebnisse sprechen *gegen* den möglichen Effekt der Fragereihenfolge „Furcht-Opfer" im SWB.[19]

Abbildung 2: Anteilswerte und deren 95%-Konfidenzintervalle in zwei Experimenten mit Veränderung der Fragereihenfolge „Opfer-Furcht" vs. „Furcht-Opfer"



Obwohl die – gemessen an Lehrbuchstandards, nicht hingegegen im Vergleich zu bei Kognitionsexperimenten üblichen Standards – relativ geringe Power dieser Tests keine endgültige Entscheidung über diese Hypothese erlaubt, gibt es bislang keine Daten, die sie stützen würden.

Damit scheinen beide Erhebungsinstrumenthypothesen – zumindest auf der Basis der vorliegenden Daten – keine ausreichende Erklärung für die Diskrepanz der Surveys zu liefern.

Prozent (MTU). Diese Anteile unterscheiden sich nicht signifikant von den Opferraten in diesen Experimenten.

19 Man könnte einwenden, dass sich der Effekt der Fragereihenfolge nicht in einer studentischen Population zeigt, wohl aber bei nichtstudentischen Populationen. Ohne vorherige Spezifikation des Mechanismus und der Vorhersage des Vorzeichens des erwarteten Effekts ist ein solches Argument wenig hilfreich. Ohnehin kann nur eine empirische Untersuchung an der allgemeinen Bevölkerung diese Frage klären. Eine solche Studie wird derzeit vorbereitet.

6. Ausschöpfung der Stichprobe

a) Bearbeitung der Sampling-Points. Selbstverständlich sollte jeder ausgewählte Sampling-Point einer Stichprobe auch bearbeitet werden. In der BRD fallen bei einem ADM-Design im Mittel 1,7 Prozent der Adressen durch Nichtbearbeitung aus (vgl. Schnell 1997: 72). In der Regel fallen diese nicht bearbeiteten Adressen meist pointweise aus, da die Nichtbearbeitungen durch Krankheit, Urlaub, Beendigung des Arbeitsverhältnisses und Verweigerung durch den Interviewer bedingt werden. In der MTU wurden von den 2100 ausgewählten Points nur 1912 bearbeitet. Im SWB wurden von 630 ausgewählten Points 603 bearbeitet. Durch nicht bearbeitete Sampling-Points fielen im SWB 6,3 Prozent der Adressen der Ausgangsstichprobe aus.

Diese hohe Ausfallrate ganzer Sampling-Points wird in der Methodendokumentation nur ungenügend begründet. Dort wird darauf verwiesen, dass vor allem in den östlichen Bundesländern ganze Straßenzüge ausfallen, weil sie Abbruch- und Sanierungsgebiet geworden sind (Methodendokumentation 1998: 7). Dies steht aber im Gegensatz zu den Ergebnissen der Feldarbeit: Die durch Nichtbearbeitung resultierende Ausfallrate der Adressen im SWB liegt mit 3,8 Prozent in den östlichen Bundesländern knapp *unter* den 3,9 Prozent in den westlichen Bundesländern. In Hinsicht auf die Ausfälle der Sampling-Points liegt auch hier der Osten mit 3,8 Prozent unter den 4,5 Prozent im Westen. Noch deutlicher wird dies bei der MTU: In den westlichen Bundesländern fielen 9,4 Prozent der Points vollständig aus, in den östlichen Bundesländern 6,9 Prozent. Die Erklärung des Instituts für diese Ausfälle lässt sich angesichts dieser Daten kaum halten.

Das durchführende Institut lag bei anderen Studien meistens deutlich über dem durchschnittlichen Anteil nichtbearbeiteter Sampling-Points aller Institute (vgl. Schnell 1997: 72). Da bei diesen Studien das gleiche Stichprobensystem auch von anderen Instituten verwendet wurde, muss die Ursache für diese hohen Anteile in den nicht dokumentierten Details der Feldarbeit liegen. Die Hauptursache für diese Besonderheit dürfte in der mangelnden Kontrolle der Tätigkeit der Interviewer zu finden sein: Die Supervisoren bemerken die Untätigkeit eines Interviewers zu spät, um noch korrigierend eingreifen zu können. Die Untätigkeit eines Interviewers an einem Sampling-Point mag zwar gelegentlich durch Urlaub oder Krankheit bedingt sein, allerdings gibt es Hinweise, dass Interviewer systematisch bestimmte ihnen gefährlich oder unzugänglich erscheinende Sampling-Points meiden (vgl. Schnell 1997: 177). In Sampling-Points mit hoher Prävalenz könnte dieses Vermeidungsverhalten der Interviewer wahrscheinlicher sein als in Sampling-Points mit niedriger Prävalenz. Hierdurch würde die Viktimisierungsrate in der resultierenden Stichprobe eine Unterschätzung der tatsächlichen Opferrate zur Folge haben. Der vollständige Ausfall eines Sampling-Points kann also systematische Zusammenhänge zum Untersuchungsthema schon allein über diesen Mechanismus besitzen.

Das durchführende Institut differenziert in der Methodendokumentation nicht zwischen nichtbearbeiteten Sampling-Points und Sampling-Points, die auf Grund des Ausfalls aller Einzeladressen trotz Bearbeitung des Sampling-Points vollständig ausfallen. Die Wahrscheinlichkeit dafür, dass ein Sampling-Point vollständig ausfällt, obwohl er vom Interviewer bearbeitet wurde, ist sehr gering: Nimmt man als Schätzung der indi-

viduellen Ausfallwahrscheinlichkeit den Anteil der Ausfälle bei den bearbeiteten Adressen (0.3), dann ergibt sich bei 8 Zielpersonen eine Wahrscheinlichkeit von 0.3^8, also gleich 0.00006561 dafür, dass alle 8 Zielpersonen ausfallen. Offensichtlich sind entweder die individuellen Ausfallwahrscheinlichkeiten deutlich höher oder die Ausfälle sind nicht unabhängig voneinander.

Nimmt man an, dass der Anteil von 1,7 Prozent nichtbearbeiteter Adressen bei anderen Instituten einen Wert darstellt, der unter den gegebenen Bedingungen der Feldarbeit in der BRD praktisch unvermeidlich ist, so liegt in dieser Studie der Anteil um ca. 2,1 Prozentpunkte (3,8–1,7) zu hoch. Eine Erklärung für diesen erhöhten Anteil kann durch die Annahme einer deutlich höheren individuellen Ausfallwahrscheinlichkeit erreicht werden. So liefert eine Ausfallwahrscheinlichkeit von 0.617 bei 8 Personen eine Wahrscheinlichkeit von 2,1 Prozent für den kompletten Ausfall eines Sampling-Points.

Eine solche Verdoppelung der Ausfallwahrscheinlichkeit kann zum Beispiel durch eine drastische Beschränkung der Kontaktversuche durch die Interviewer bedingt sein. Hierdurch würden die leichter erreichbaren Zielpersonen in der Stichprobe überrepräsentiert. Sollten schwerer erreichbare Zielpersonen höhere Viktimisierungswahrscheinlichkeiten aufweisen, wird durch diesen Mechanismus ebenfalls eine Unterschätzung der Opferwerdung in der Stichprobe erfolgen.

Eine solche Verdoppelung der Ausfallwahrscheinlichkeiten kann aber auch durch erhöhte Verweigerungsquoten der Zielpersonen bedingt sein.[20] Sollten Zielpersonen mit Opfererfahrung höhere Verweigerungsquoten aufweisen als Zielpersonen ohne Opfererfahrung, würde die Opferrate in der Stichprobe ebenfalls unterschätzt.

Damit besteht die Möglichkeit, dass drei verschiedene systematisch mit der Opferrate zusammenhängende Mechanismen zu Ausfällen ganzer Sampling-Points führen können. Gerade in Viktimisierungssurveys können solche oft fälschlich als „qualitätsneutrale Ausfälle" bezeichneten Ausfälle zur deutlichen Unterschätzung der Viktimisierungsrate führen (Schnell 1997: 72). Eine direkte Prüfung, ob diese möglichen Mechanismen tatsächlich vorliegen oder nicht, ist mit den vorliegenden Daten nicht möglich.[21] Trotzdem kann die entscheidende Frage, ob das Ausmaß der Ausfälle durch Nichtbearbeitung bei Annahme systematischer Ausfallmechanismen ausreicht, um den Unterschied zwischen den Untersuchungen zu erklären, beantwortet werden.

Nimmt man an, dass die wegen Nichtbearbeitung ausgefallenen Adressen den gleichen Anteil von Ausfällen erlitten hätten wie die anderen Adressen, so ergibt eine einfache Rechnung, dass erst eine Opferrate von 45 Prozent in diesen Sampling-Points zu

20 Couper und Groves (1996: 185) stellten für sechs landesweite Surveys der USA geringere Kooperationsbereitschaft in Gebieten mit hoher Kriminalitätsrate fest. Sie vermuten, dass eine Messung der Kriminalitätsfurcht die Effekte der Kriminalitätsraten auf die Kooperationsbereitschaft erklären würde (Couper und Groves 1996: 186).

21 Das durchführende Institut untersuchte zur Überprüfung systematischer Ausfälle von Sampling-Points die Streuung der Ausfälle über Bundesländer und Gemeindegrößenklassen (vgl. Methodendokumentation 1998: 6). Da die oben erläuterten möglichen Mechanismen für den Ausfall ganzer Sampling-Points sich aber keineswegs zwischen den Bundesländern oder den Gemeindegrößenklassen unterschiedlich auswirken müssen, kann diese Art von Test als bedeutungslos betrachtet werden.

einer Überlappung der Konfidenzintervalle der MTU mit den unkorrigierten Konfidenzintervallen des SWB geführt hätte.[22] Die Ausfälle durch Nichtbearbeitung reichen allein also kaum aus, um den Unterschied zwischen den Surveys zu erklären. Allenfalls kann ein kleiner Teil des Unterschieds zwischen den Surveys durch diese Unterschiede in der Feldarbeit erklärt werden.

b) Ausschöpfungsrate. Fast immer wird eine global berechnete Ausschöpfungsrate fälschlich als Qualitätskriterium für einen Survey verwendet. Entsprechend liegt die Hypothese nahe, dass die Unterschiede zwischen den Surveys durch unterschiedliche Ausschöpfungen bedingt sein könnten. Dies ist allerdings falsch: Im SWB lag die Ausschöpfungsrate bei 67,6 Prozent im Westen bzw. 69,1 Prozent im Osten, für den MTU werden 66,3 bzw. 69,4 Prozent berichtet (Methodendokumentation 1998: 910).

IV. Erklärungen durch üblicherweise nicht berichtete Details der Feldarbeit

Da die bisher erörterten üblichen Erklärungsmechanismen in diesem Fall keine plausiblen Ursachen für die beobachteten Unterschiede liefern konnten, wurde nach weiteren möglichen Ursachen gesucht. Hierbei zeigt sich rasch, dass Informationen über Details der Feldarbeit benötigt werden, die üblicherweise weder Bestandteil der Datensätze noch der Erhebungsberichte sind. Hierzu gehören vor allem Angaben über das tatsächliche Verhalten der Interviewer. Die bei kommerziellen Instituten in der BRD übliche Praxis der Datenerhebung führt dazu, dass der Interviewer große Spielräume bei der Auswahl der Zielperson und dem Interviewablauf besitzt. Durch das Verhalten der Interviewer bei der Auswahl der Zielhaushalte, bei der Kontaktierung der Zielpersonen und schließlich bei der Befragung selbst kann es zu systematischen Ausfällen vollständiger Interviews oder von Teilen des Interviews kommen. Das Verhalten der Interviewer hängt maßgeblich von den organisatorischen Rahmenbedingungen ab: Interviewerentlohnungsystem, Interviewerschulung, Vorgaben bezüglich der Auswahl der Zielperson und des Interviewablaufs sowie die Kontrolle dieser Vorgaben durch Supervisoren. Sollten sich diese Faktoren zwischen den Erhebungen unterscheiden, wären die Unterschiede zwischen den Erhebungen durch diese in der Regel undokumentierten Details der Feldarbeit erklärbar.

---

22 Wenn r die Responsequote bei den bearbeiteten Adressen, n_u die Anzahl der unbearbeiteten Adressen, n_b die Anzahl der bearbeiteten Adressen, o_b der Anteil der Opfer bei den bearbeiteten Adressen und o_u der Anteil der Opfer bei den unbearbeiteten Adressen ist, dann ergibt sich die zu erwartende Zahl der durch Bearbeitung zusätzlich erhaltenen Fälle mit

$$n = r \cdot n_u$$

Nimmt man an, dass für diese n Fälle der Anteil der Opfer o_u um den Faktor B höher liegt als bei den bearbeiteten Adressen, so ergibt sich die Opferrate im gesamten Datensatz durch

$$o = \frac{B \cdot o_u \cdot n + o_b \cdot n_b}{n_b + n}$$

1. Entlohnungssystem der Interviewer

Die Einzelheiten des Interviewerentlohnungssystems des durchführenden Instituts sind nicht publiziert. Es ist aber unwahrscheinlich, dass das Interviewerentlohnungssystem zwischen den Studien variiert. Effekte unterschiedlicher Entlohnungssysteme werden deshalb hier nicht weiter als mögliche Ursachen untersucht.

2. Interviewerschulung

Das durchführende Institut verfügt bundesweit über einen Interviewerstamm von ca. 2.750 freiberuflichen Interviewern. Nach eigenen Angaben befinden sich darunter rund 500 Interviewer mit spezieller Ausbildung für Forschungsprojekte mit besonderem thematischen Schwierigkeitsgrad. Es wäre denkbar, dass die Interviewer des SWB zu diesem besser geschulten Interviewerstamm gehören. Zwar teilte das Institut auf Anfrage mit, dass vermutlich keine systematischen Differenzen zwischen den Stäben dieser beiden Surveys bestehen, allerdings ist eine systematische Prüfung dieser Hypothese nicht mehr möglich. Eine zumindest partielle Prüfung dieser Hypothese wäre durch einen Vergleich der Interviewernummern in beiden Untersuchungen möglich. Leider ist das Nummerierungssystem beider Untersuchungen unterschiedlich, sodass auch dieser Vergleich nicht durchgeführt werden konnte. Allerdings legt die Literatur zum Interviewerverhalten nahe, dass der Effekt der Interviewerschulung gegenüber dem Effekt der Interviewerkontrolle eher gering ist (Fowler und Mangione 1990: 131).

3. Auswahl der Zielperson

Die beiden Stichproben unterscheiden sich stark in Bezug auf die Intervieweranweisungen zur Auswahl der Zielhaushalte und Zielpersonen. Im SWB wurde den Interviewern ein Startpunkt zugewiesen und ausgehend von diesem Startpunkt nach streng geregeltem „Random-Route" jede dritte Adresse notiert bis zu einer Gesamtzahl von 23 Privathaushalten. Das Institut markierte vorab acht bzw. neun Adressenfelder pro Point (Methodendokumentation 1998: 4). Die Adressen, die von den Interviewern in die markierten Felder eingetragen wurden, bildeten die Brutto-Ausgangsstichprobe der zu kontaktierenden Haushalte. Den Interviewern wurde im SWB mindestens drei Kontaktaufnahmeversuche vorgeschrieben. Die Zahl der bis zum erfolgten Interview notwendigen Kontaktversuche musste vom Interviewer in den Fragebogen eingetragen werden.[23] In der Methodendokumentation heißt es, dass laut Angaben auf den Kontaktprotokollen (zumindest in den städtischen Gebieten) die Zahl der vorgeschriebenen Kontaktversuche in den meisten Fällen überschritten wurde.

In der MTU erfolgte die Auswahl innerhalb der Sampling-Points nach „allgemein üblichen Random-Route-Auswahlmodalitäten" (Methodendokumentation 1998: 5). In

23 Leider sind diese Angaben aber nicht Bestandteil des Datensatzes. Auf Anfrage erklärte das Institut, dass diese Angaben nicht maschinell erfasst wurden.

der Methodendokumentation heißt es „Ausgehend von dieser Startadresse hatten sie nach einem vorgegebenen Wegeschema jeden dritten Haushalt zu kontaktieren ...". Falls es sich dabei um einen Zielhaushalt der definierten Grundgesamtheit handelt, „hatten sie sich um eine Realisierung eines Interviews zu bemühen". Innerhalb der vorgegebenen Befragungsroute sollten die Interviewer nach Möglichkeit zehn Interviews realisieren. Eine Kontrolle der Zielhaushalte wurde nicht durchgeführt. Für die MTU Stichprobe wurden von den Interviewern keine Kontaktprotokolle geführt. Auch eine Spezifizierung der Ausfallgründe wurde nicht vorgenommen. Die Zahl der Kontaktversuche kann vom Institut nicht überprüft werden.

Die Auswahlvorgaben in der MTU legen es nahe, dass die Interviewer versuchten, die Minimalforderung von zehn Interviews pro Point zu erfüllen. Es kann vermutet werden, dass die Interviewer im MTU bei misslungenen Kontaktierungsversuchen des laut Random-Route zu erreichenden Zielhaushalts diesen eher durch einen anderen Zielhaushalt ersetzten. Ebenso dürften die Interviewer im MTU verweigernde Zielpersonen bzw. Zielhaushalte eher durch kooperativere Haushalte ersetzt haben.

Wie bereits oben erwähnt, gibt es aus anderen Untersuchungen empirische Hinweise auf erhöhte Viktimisierungsraten bei Schwererreichbaren bzw. Verweigerern. Sollte sich diese Tendenz auch in diesen Untersuchungen durchsetzen, so muss das zu erwartende Interviewerverhalten im MTU zu einem deutlich niedrigeren Anteil der berichteten Opferwerdung als im SWB führen.

a) Gewichtung als Indikator für Ausfallprobleme. Man könnte versuchen, die vom Institut berechnete Gewichtung als Indikator für solche speziellen Ausfallprobleme im MTU und SWB zu verwenden. Dies wäre dann gegeben, wenn sich auffällig hohe Gewichte in einigen Sampling-Points bzw. bei einigen Interviewern häufiger ergäben als in anderen Sampling-Points bzw. bei anderen Interviewern.[24] Zwar finden sich im MTU tatsächlich einige auffällig hohe Gewichte, diese stammen aber aus unterschiedlichen Points bzw. von unterschiedlichen Interviewern. Die Verteilung der Gewichtungsvariablen liefert damit keine Hinweise auf besondere Ausfallprobleme.[25]

b) Modellierung der Ausfälle. Die Datensätze bieten selbst keine Möglichkeit, den Einfluss selektiver Ausfälle auf die Ergebnisse zu ermitteln. Trotzdem können solche Ef-

24 In die Gewichtung wurden vom Institut die Merkmale Bundesland, politische Gemeindegrößenklasse, Geschlecht und Altersverteilung der Grundgesamtheit einbezogen (Methodendokumentation 1998: 12).

25 Ein Hinweis auf Selektionseffekte durch die Interviewer könnte der Vergleich der Sampling-Points im MTU und SWB bieten. So könnte man beim Vergleich gleicher Points feststellen, ob sich die Gewichtungsvariablen zwischen den beiden Studien für diesen Point auffällig unterscheiden. Sollten die Ausfälle durch den Interviewer unsystematisch erfolgen, dann sollten innerhalb eines Points nicht nur die inhaltlichen Variablen, sondern auch die erhebungstechnischen Variablen wie die Gewichtung in beiden Studien ähnliche Werte besitzen. Dieser Vergleich lässt sich jedoch mit den vorliegenden Datensätzen nicht führen, da auch das Nummerierungssystem der Points verändert wurde. Der Versuch, die Points nachträglich durch Kombinationen von Regierungsbezirk und BIK-Gemeindegrößenklassen zuzuordnen, führte zu keiner eindeutigen Zuordnung und konnte daher nicht weiter verfolgt werden.

fekte auf verschiedene Art und Weise abgeschätzt werden. Würden sich Respondenten und Nonrespondenten systematisch in Hinsicht auf die Opferwerdung unterscheiden, so lässt sich entsprechend den oben durchgeführten Berechnungen auch hier abschätzen, wie groß die Opferrate bei den ausgefallenen Zielpersonen sein müsste, um zu einer Überlappung der Konfidenzintervalle zu gelangen. Die Berechnungen ergeben, dass bei einer Opferrate von 22 Prozent bei den Nonrespondenten sich die Konfidenzintervalle der beiden Untersuchungen überlappen. Dies entspräche einer 1.39-fachen Opferrate bei den Nonrespondenten.

Verwendet man um Designeffekte korrigierte Konfidenzintervalle, so würden sich die Konfidenzintervalle vermutlich bereits bei einer Opferrate von 20 Prozent, also dem 1.26-fachen der Opferrate der Respondenten, überlappen. Ein solcher Unterschied zwischen Respondenten und Nonrespondenten erscheint zunächst nicht sofort als unmöglich.

Der Mechanismus des Einflusses selektiver Ausfälle auf die Ergebnisse lässt sich auch direkt modellieren. Es scheint plausibel anzunehmen, dass die Wahrscheinlichkeit, Opfer zu werden, umso höher ist, je häufiger das Haus verlassen wird. Dies gilt zumal dann, wenn wie in diesem Fall die erfragten Delikte überwiegend Delikte sind, die außerhalb der eigenen Wohnung stattfinden. Weiterhin sind Personen, die häufiger außer Haus sind, auch schwerer erreichbar. Damit erscheint die Annahme einer positiven Korrelation zwischen der Zahl der Kontaktversuche, die notwendig sind, um die Zielperson zu erreichen, und der Wahrscheinlichkeit, Opfer zu werden, ebenfalls plausibel.[26]

Zur Überprüfung des vermuteten Zusammenhangs zwischen der Zahl der Kontaktversuche bis zum Interview bzw. dem Ausmaß der außerhäuslichen Aktivitäten und der berichteten Opferwerdung stehen kaum empirische Daten zur Verfügung, da nur zwei der uns zugänglichen Studien entsprechende Indikatoren enthalten: Die Studie „Soziale Infrastruktur 1976/1977" (ZA Nr.1098/1099) und der „British Crime Survey" von 1992 (ICPSR Nr. 06717).

Für die Studie „Soziale Infrastruktur" ergibt sich für beide Wellen zwischen der Zahl der Kontaktversuche und Opferwerdung eine signifikante Korrelation von $r = .07$ ($p = .05/p = .01$). Berechnet man aus acht Variablen zur Freizeitverwendung einen Aktivitätsniveau-Index, so ergibt sich zwischen Opferwerdung und Aktivitätsniveau in beiden Wellen eine Korrelation von $r = .13$ ($p = .00$). Die Daten des „British Crime Surveys" zeigen ebenfalls eine signifikante Korrelation von $r = .05$ ($p = .00$) zwischen Opferwerdung bei Gewaltdelikten und der Zahl der Kontaktversuche.[27] Der Zusammenhang zwischen Opferwerdung und der Zahl der Kontakte bis zum Interview kann daher als zwar schwach, aber nachgewiesen angesehen werden.

---

26 Ebenso liegt eine Korrelation zwischen Kooperationsbereitschaft und Opferwerdung nahe.

27 Eine genauere Analyse zeigt, dass der Zusammenhang zwischen Opferwerdung und Erreichbarkeit nicht linear ist: Befragte, die bereits beim ersten Kontaktversuch erreicht wurden, zeigen eine deutlich niedrigere Opferrate; Befragte, die erst nach 6 bis 12 Kontaktversuchen erreicht wurden, zeigen eine deutlich höhere Opferrate, als beim Vorliegen eines linearen Zusammenhangs zu erwarten wäre.

Da die zum Nachweis des Zusammenhangs zwischen Zahl der Kontaktversuche bis zum Interview und Opferwerdung herangezogenen Surveys selbst unter Ausfällen leiden, sollte die beobachtete Korrelation um den Effekt der Ausfälle korrigiert werden. Für die deutsche Studie wurde keine Ausschöpfung berichtet, für den „British Crime Survey" liegt sie bei 76,7 Prozent. Im SWB lag die Ausschöpfungsrate bei 68,1 Prozent, in der MTU bei 66,9 Prozent. Im Folgenden wird daher von 70 Prozent Ausschöpfung ausgegangen. Nimmt man an, dass die Ausfälle ausschließlich durch Nichterreichbarkeit verursacht werden, dann folgt aus den Formeln zum „Restriction of Range"-Problem bei einer beobachteten Korrelation von ca. 0.125 eine Korrelation von 0.2 in der Population.[28] Eine Korrelation von 0.125 in der realisierten Stichprobe erscheint für die heutige BRD unter Berücksichtigung auch anderer als Gewaltdelikte zunächst nicht unplausibel, sodass im Folgenden von einer tatsächlichen Korrelation von 0.2 ausgegangen wird.

Bezeichnet man mit p den Anteil der am schwersten zu erreichenden Personen einer Stichprobe und nimmt an, dass die Erreichbarkeit C standardnormalverteilt ist, dann entspricht der Schwellenwert T der Erreichbarkeit, jenseits dessen niemand mehr erreicht wird, der inversen Verteilungsfunktion der Standardnormalverteilung (F^{-1})

$$T = F^{-1}(1 - p)$$

Bei 30 Prozent Ausfällen entspricht dies einem Wert von T = .5244. Durch die Ausfälle verändert sich der Mittelwert der beobachteten Kontaktzahlen. Für standardnormalverteilte Variablen lässt sich der Erwartungswert $E(x)$ bei einem Schwellenwert $T > 0$ unter Verwendung der Dichtefunktion der Standardnormalverteilung Φ berechnen mit

$$E(X) = \frac{-\Phi(T)}{F(T)}$$

(Rydberg 1963: 52–53). Bei 30 Prozent Ausfällen liegt der Erwartungswert standardnormalverteilter Erreichbarkeit also bei –.348/.7 = –.497.

Der Erwartungswert der in der unvollständig gezogenen Stichprobe beobachteten Opferrate V hängt von der Stärke der Korrelation r_{VC} in der Grundgesamtheit zwischen der Opferrate und der Erreichbarkeit und dem Erwartungswert der Erreichbarkeit $E(C)$ in der unvollständig gezogenen Stichprobe ab. Für standardnormalverteilte Variablen gilt:

$$E(V) = r_{VC} \cdot E(C)$$

---

28 Eine einfache Korrekturformel für Korrelationen, für die lediglich der Anteil der ausgefallenen Stichprobe bekannt sein muss, findet sich bei Sands et al. (1978: 748). In der weiter unten erläuterten Notation lautet sie:

$$R = \frac{r}{\sqrt{(1-r^2)\cdot\left[1+\frac{T\cdot\Phi(T)}{F(T)}-\left(\frac{\Phi(T)}{F(T)}\right)^2\right]+r^2}},$$

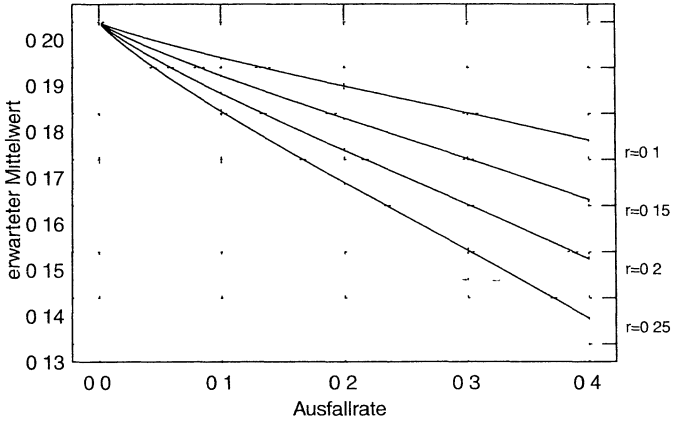
wobei R die korrigierte Korrelation in der Population, r die beobachtete Korrelation und T der Schwellenwert ist.

(Johnson und Kotz 1972: 112). Bei einer angenommenen Korrelation von 0.2 und 30 Prozent Ausfallen erhalt man also

$$E(V) = 0.2 \quad (-0.497) = -0.0994.$$

Da die Opferwerdung aber einen Mittelwert von 0.2 und eine Standardabweichung von 0.4 besitzt, muss der Erwartungswert noch einer umgekehrten z-Transformation unterzogen werden, also mit der Standardabweichung multipliziert und der Mittelwert addiert werden. Damit ergibt sich bei 30 Prozent Ausfallen der am schwersten zu erreichenden Personen eine erwartete Opferwerdung von (–0.0994 0.4)+0.2 = 0.16. Dieser geschatzte Anteil entspricht der beobachteten Opferrate in der MTU sehr gut. Die *Abbildung 3* zeigt ein den Formeln entsprechendes Nomogramm fur den Zusammenhang zwischen Opferrate und Ausfallen für unterschiedliche Korrelationen zwischen Erreichbarkeit und Opferwerdung.

Abbildung 3· Nomogramm des Zusammenhangs zwischen Erreichbarkeit und Anteil der Opferwerdung fur verschiedene angenommene Korrelationen



Diese Berechnungen basieren auf einigen statistischen Annahmen, wie z.B. normalverteilten Variablen. Im gegebenen Fall sind diese Annahmen nicht erfullt. Um zu prüfen, ob das berichtete Ergebnis sich auch bei Verletzung der Annahmen zeigt, wurde eine Simulationsstudie durchgefuhrt.

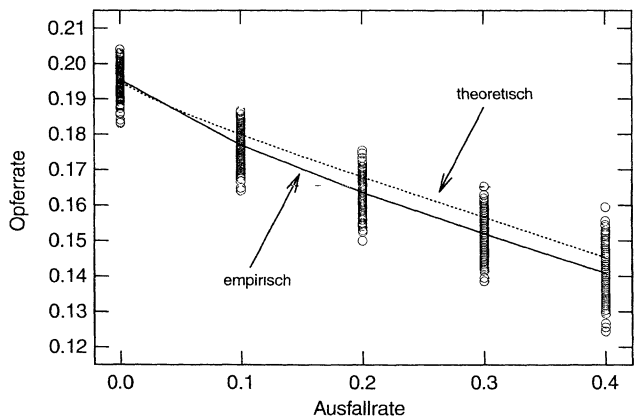
Die Simulation basiert auf einem zweifaktoriellen Design. Variiert wurde die angenommene Korrelation zwischen Opferwerdung und Erreichbarkeit (0.25 bis 0.4 in 0.05 Schritten) sowie der Anteil der durch unterschiedliche Erreichbarkeit ausgefallenen Personen (10–40 Prozent in 10 Prozent-Schritten).[29] Abhangige Variable war der

29 Die Variablen wurden durch einen Generator, der multivariat normalverteilte Variablen mit vorgegebener Korrelationsstruktur uber eine Cholesky-Dekomposition ergibt, erzeugt. Die Simulation wurde mit *Stata* durchgefuhrt Die vorgegebenen Korrelationen wurden hoher gewahlt als die empirisch erwarteten, da sich durch die unten beschriebene notwendige Dichotomisierung und Transformation eine Verringerung der beobachteten Korrelationen ergibt

Mittelwert der Opferwerdung in der verbleibenden Stichprobe. Jede der 4 Faktorstufen wurde 500-mal wiederholt. Bei jeder Wiederholung wurden für 3000 Personen je zwei Variablen erzeugt. Nach der Erzeugung wurde eine der Variablen entsprechend dem Anteil der Opferwerdung (19,5 Prozent) dichotomisiert. Die andere Variable wurde so transformiert, dass sie einer empirischen Kontaktzahlverteilung entspricht.[30] Aus jedem der so erzeugten Datensätze wurden entsprechend der simulierten Erreichbarkeit jeweils 10–40 Prozent gelöscht.

Wie oben bereits begründet, ist hier vor allem das Ergebnis für eine angenommene Korrelation von 0.2 zwischen Erreichbarkeit und Opferwerdung in der Grundgesamtheit von Interesse. Wie die *Abbildung 4* zeigt, ergibt ein Vergleich der simulierten Opferraten mit den theoretisch erwarteten Opferraten eine ausgezeichnete Übereinstimmung zwischen Simulation und den Berechnungen durch die Selektionsgleichung.[31]

Abbildung 4: Theoretisch vorhergesagte und simulierte Anteile der Opferwerdung in Abhängigkeit vom Anteil der Ausfälle



Sowohl die Ersetzung der Nichtrespondenten durch Personen mit angenommener höherer Opferwerdung, als auch die Modellierung mit einer Selektionsgleichung, als auch die Simulation des Ausfallprozesses führen zum gleichen Ergebnis: Sollten die Schwererreichbaren und Verweigerer über höhere Opferwerdungen verfügen bzw. sollte es eine – auch nur schwache – Korrelation zwischen Opferwerdung und Erreichbarkeit

30 Die Variable wurde einer Powertransformation mit einem Exponenten von 3.0 unterzogen und anschließend in 14 Stufen kategorisiert, wobei die Quantile den Kontakthäufigkeiten im ALLBUS 1996 entsprechen. Durch diese Transformationen verringern sich die Korrelationen bei einem Teil der Wiederholungen beträchtlich. Für die Analysen wurden diejenigen Wiederholungen ausgewählt, bei denen die realisierten Korrelationskoeffizienten zwischen der dichotimisierten Opferwerdung und der transformierten und kategorisierten Erreichbarkeit nicht mehr als 5 Prozent vom vorgegebenen Wert abwichen. In dem im Folgenden beschriebenen Beispiel reduziert sich die Zahl der Wiederholungen der Zelle damit auf jeweils 275.

31 Die durchgezogene Linie in der Abbildung ist eine nichtparametrische Regression (lowess mit f=0.5).

geben, dann reicht dies zur Erklärung der Unterschiede zwischen den Surveys vollkommen aus. Damit wäre ein möglicher Mechanismus, auf den die Unterschiede zurückgehen könnten, aufgezeigt. Eine nachträgliche Korrektur der Stichproben wird dadurch aber nicht möglich.[32]

4. Arbeitsbelastung, Interviewdauer und Teilfälschungen

Zusätzlich zu den Fehlern bei der Auswahl der Zielhaushalte und Zielpersonen kann mit systematischen Fehlern im Antwortprozess gerechnet werden. Insbesondere in Studien mit hoher Interviewerbelastung ist es denkbar, dass Interviewer die Befragungszeit dadurch abkürzen, dass sie den Fragebogen nur unvollständig erheben. Dies bietet sich vor allem bei Filterfragen oder Fragebatterien an.[33] Solche Bedingungen liegen bei diesen Untersuchungen vor: Die Arbeitsbelastung in der MTU lag deutlich höher als im SWB, die Kontrollen waren in der MTU nicht so streng wie im SWB und der Fragebogen ähnelt einer Fragebatterie, die durch eine Filterfrage vermieden werden kann. Die Arbeitsbelastung pro Interviewer war in der MTU ungewöhnlich hoch: In der MTU wurden von jedem Interviewer im Mittel 22 Interviews durchgeführt (vgl. *Abbildung 5*), das Maximum liegt bei 49 Interviews. Im SWB haben die 413 Interviewer durchschnittlich 8 Interviews durchgeführt.[34] Zwanzig Interviewer (das sind etwa 5 Prozent der Interviewer des SWB) führten 25 oder mehr Interviews durch, das Maximum liegt hier bei 53 Interviews.[35]

Sollten die Interviewer verschiedener Studien in unterschiedlichem Ausmaß dazu neigen, die Interviews abzukürzen, wäre ein weiterer Mechanismus, der die Unterschiede zwischen den Surveys bedingen kann, gefunden. Dieser Effekt soll im Folgenden abgeschätzt werden.

Angenommen, die Interviewer des SWB führten auf Grund der strengeren Kontrolle der Feldarbeit ihre Interviews gemäß der Intervieweranweisungen durch und ein be-

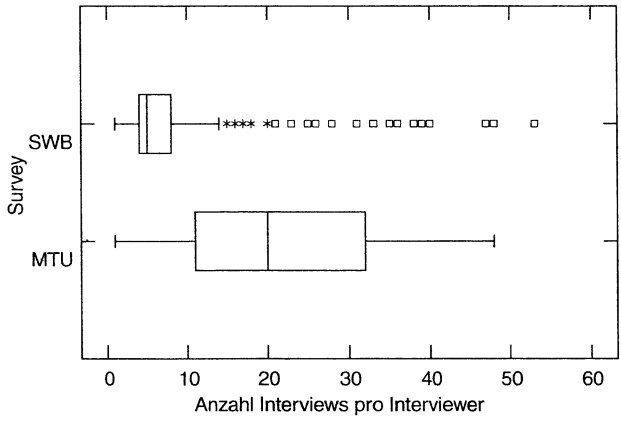
32 Im Rahmen der Analysen wurde versucht, die vermutete Selektivität der Ausfälle durch ein „Propensity-Modell" zu korrigieren (Schnell 1997). Hierzu wurde mit Hilfe eines Datensatzes, in dem die Zahl der bis zum Interview notwendigen Kontaktversuche protokolliert ist, die Erreichbarkeitswahrscheinlichkeit für Personen mit bestimmten Merkmalskombinationen geschätzt. Anschließend wurden in der MTU und im SWB die Personen, die die entsprechenden Merkmalskombinationen aufweisen, mit der Inversen dieser Erreichbarkeitswahrscheinlichkeit gewichtet. Dadurch sollte der systematische Selektionseffekt zurückgehen und sich die Opferraten in SWB und MTU angleichen. Tatsächlich zeigt sich ein leichter Anstieg der Opferraten in beiden Studien. Dies deutet auf die Richtigkeit der Annahme über eine Korrelation zwischen der Erreichbarkeit und der Opferwahrscheinlichkeit hin. Die Korrelation zwischen geschätzter Erreichbarkeit und Opferstatus liegt bei r=–.12. Die durch das Propensity-Modell erhoffte Anpassung der Opferraten zwischen MTU und SWB ließ sich nicht beobachten. Dies könnte sowohl auf die schlechte Anpassung des Propensity-Modells als auch auf deutliche Selektionseffekte in beiden Surveys hindeuten.

33 Eine Diskussion solcher Effekte findet sich bei Fowler und Mangione (1990).

34 Diese Zahl basiert auf den Angaben im Datensatz; der Feldbericht weist demgegenüber nur 397 Interviewer für den SWB aus (Methodendokumentation 1998: 10).

35 Dieser Interviewer führte in 8 Sampling-Points insgesamt 53 Interviews durch.

Abbildung 5: Arbeitsbelastung der Interviewer in den beiden Surveys



stimmter Prozentsatz der Interviewer der MTU verkürzte das Interview dadurch, dass er unabhängig von der tatsächlichen Opfererfahrung der Befragten „keine Viktimisierung" eintrug, um die darauf folgenden Fragen zu umgehen.

Der Effekt dieses Verhaltens wäre überraschend stark. Geht man davon aus, dass in 10 Prozent aller Interviews des MTU die Befragten unabhängig von ihrem tatsächlichen Opferstatus vom Interviewer als Nichtopfer klassifiziert wurden, dann kann man den Effekt dieses Verhaltens mit zwei verschiedenen Methoden berechnen: Analytisch oder über eine Simulation. Simuliert man dieses angenommene Verhalten der Interviewer, so ergeben sich in 37 von 100 simulierten Datensätzen Opferraten, deren Konfidenzintervalle sich mit denen der MTU überlappen.[36] Analytisch ergibt sich bei einem angenommenen wahren Wert von 19,5 Prozent Opfer und 10 Prozent falsch klassifizierter Fälle eine minimale Opferrate von 9,5 Prozent; im Mittel ergäbe sich eine Opferrate von ca. 17,6 Prozent.[37] Neben den selektiven Ausfällen besteht also in der Art der Interviewdurchführung bzw. den Interviewteilfälschungen ein weiterer Mechanismus, der den Unterschied zwischen den Untersuchungen erklären könnte.

36 Mit Stata wurden, ausgehend bei einem n von 3272 und einem Opferanteil von 19,5 Prozent mit 100 Wiederholungen, jeweils 10 Prozent zufällig ausgewählter Fälle als „Nicht-Opfer" klassifiziert und die entsprechenden Konfidenzintervalle für die sich dann ergebenden Opferraten mit der Clopper-Pearson-Methode berechnet. Gezählt wurde die Zahl der simulierten Datensätze, bei denen die untere Grenze des Konfidenzintervalls unterhalb der oberen Grenze des Konfidenzintervalls der MTU lag.

37 Im schlimmst möglichen Fall sind von den angenommenen 10 Prozent abgekürzten Interviews alle tatsächlich Opfer. Also verringert sich der Anteil von 19,5 auf 9,5 Prozent. In der Regel dürfte die fälschliche Klassifikation als Nicht-Opfer unabhängig vom tatsächlichen Opferstatus sein. In diesem Fall würden also 1,95 Prozent der Opfer als Nicht-Opfer klassifiziert. Also verringert sich entsprechend der Anteil der Opfer von 19,5 auf 17,55 Prozent.

V. Zusammenfassende Diskussion und Konsequenzen

Die Berücksichtigung der Design-Effekte bei den Konfidenzintervallen zeigt, dass der Unterschied zwischen den Untersuchungen kleiner sein könnte, als es zunächst scheinen mag. Trotzdem bleiben deutliche Unterschiede zwischen beiden Untersuchungen, die nicht durch Zufall zu erklären sind.

Von allen hier untersuchten Mechanismen scheinen vor allem zwei Mechanismen die beobachteten Unterschiede erklären zu können:
- Ausfälle durch Schwererreichbare mit höherer Opferrate
- Interviewerverhalten durch Abkürzen des Fragebogens

Beide Mechanismen reichen allein aus, um mit plausiblen Annahmen Unterschiede in der beobachteten Größe hervorzubringen. Beide Mechanismen wurden bislang in der Literatur nicht mit ihren möglichen quantitativen Effekten diskutiert.

Für künftige Untersuchungen muss Folgendes festgehalten werden:
- Die Konfidenzintervalle bundesweiter sozialwissenschaftlicher Surveys sind erheblich größer als bei einfachen Zufallsstichproben.
- Die erforderliche Fallzahl für zuverlässige Schätzungen ist daher weit größer als bei naiver Berechnung.
- Die Details der Feldarbeit sind von entscheidender Bedeutung und müssen daher ausführlich auch im Datensatz dokumentiert werden. Dies gilt insbesondere für so genannte „Nonresponse-Records".
- Jede Viktimisierungsstudie benötigt eine eigene Nonresponse-Studie.
- Das Interviewerverhalten muss im Detail kontrolliert werden. Dies scheint eher durch telefonische Befragungen möglich. Daher sollte für Viktimisierungsstudien die Moglichkeit einer telefonischen Befragung oder zumindest die Möglichkeit eines Mixed-Mode-Surveys durch Methodenexperimente vorab geklärt werden.

Literatur

Agresti, Alan, und *Brent A. Coull,* 1998: Approximate is better than „exact" for Interval Estimation of Binomial Proportions, The American Statistician 52, 2: 119–126.

Couper, Mick P., und *Robert M. Groves,* 1996: Social Environmental Impacts on Survey Cooperation, Quality and Quantity 30: 173–188.

Fowler, Floyd J., und *Thomas W. Mangione,* 1990: Standardized Survey Interviewing: Minimizing Interviewer Related Error. Newbury Park: Sage.

Groves, Robert M., 1989: Survey Errors and Survey Costs. New York: Wiley.

Heinz, Wolfgang, und *Gerhard Spieß,* 1995: Viktimisierung, Anzeigeerstattung und Einschätzung der Arbeit der Polizei durch die Bürger – Analysen anhand der Bevölkerungsbefragung in den Projektstädten. S. 93–122 in: *Thomas Feltes* (Hg.): Kommunale Kriminalprävention in Baden-Württemberg. Holzkirchen: Felix-Verlag.

Johnson, Norman L., und *Samuel Kotz,* 1972: Continuous Multivariate Distributions. New York: Wiley.

Kish, Leslie, 1965: Survey Sampling. New York: Wiley.

Lynn, Peter, 1997: Sampling Frame Effects on the British Crime Survey, Journal of the Royal Statistical Society, Series A, 160, 2: 253–269.

Methodendokumentation, 1998 Methodenbericht zur technischen Organisation und Durchführung des Studienprojektes „Opferumfrage 97" im Rahmen des Sozialwissenschaften-Bus III/97 und der Mehrthemengroßumfrage 1997

Rydberg, Sven, 1963 Bias in Prediction Stockholm Almquist och Wiksell

Sands, William A., Edward F Alf und *Norman M Abrahams,* 1978 Correction of Validity Coefficients for Direct Restriction in Range Occasioned by Univariate Selection, Journal of Applied Psychology 63, 6 747–750

Stata Corporation, 1997 Statistical Software, Release 5, Reference Manual – College Station

Schnell, Rainer, 1991 Wer ist das Volk? Zur faktischen Grundgesamtheit bei allgemeinen Bevölkerungsumfragen, Kölner Zeitschrift für Soziologie und Sozialpsychologie 43 106–137

Schnell, Rainer, 1997 Nonresponse in Bevölkerungsumfragen Opladen Leske + Budrich

Schnell, Rainer, Paul B Hill und *Elke Esser,* 1999 Methoden der empirischen Sozialforschung, 6 Auflage München Oldenbourg

Turner, Charles F., 1984 Why do Surveys Disagree? Some Preliminary Hypotheses and Some Disagreeable Examples S 157–214 in *Charles F Turner* und *Elizabeth Martin* (Hg.) Surveying Subjective Phenomena, Vol 1 New York Russel Sage

Korrespondenzanschrift Prof Dr Rainer Schnell, Universität Konstanz, Fakultät für Verwaltungswissenschaft, Universitätsstr 1, D-78434 Konstanz

E-Mail Rainer Schnell@uni-konstanz de